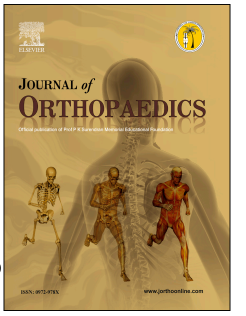
Accepted Manuscript



3D reconstruction does not improve agreement and results in an increase in surgical indications in proximal humeral fractures

Carlos Torrens, Raquel Marí, Marta Cuenca, Tulia Ferrer, Klaus Langohr P, Fernando Santana

PII: S0972-978X(18)30264-2

DOI: 10.1016/j.jor.2018.08.041

Reference: JOR 623

To appear in: *Journal of Orthopaedics*

Received Date: 12 June 2018

Accepted Date: 25 August 2018

Please cite this article as: Torrens C, Marí R, Cuenca M, Ferrer T, Langohr P K, Santana F, 3D reconstruction does not improve agreement and results in an increase in surgical indications in proximal humeral fractures, *Journal of Orthopaedics* (2018), doi: 10.1016/j.jor.2018.08.041.

Title

3D reconstruction does not improve agreement and results in an increase in surgical indications in proximal humeral fractures

Running title

3D reconstruction in proximal humeral fractures

Authors

Carlos Torrens [a], Raquel Marí [a], Marta Cuenca [a], Tulia Ferrer [a], Klaus Langohr P [b], Fernando Santana [a].

a Department of Orthopedics. Hospital del Mar. Parc de Salut Mar.

b Integrative Pharmacology and Systems Neuroscience Research Group, Neurosciences Research Program, Hospital del Mar Medical Research Institute (IMIM). Department of Statistics and Operations Research, Universitat Politècnica de Barcelona/ Barcelonatech.

Correspondence

Carlos Torrens

86925@parcdesalutmar.cat

Hospital del Mar

Passeig Marítim 25-29

08003 Barcelona

Spain

ABSTRACT

The objective of this study is to analyze the effect of different imaging modalities in treatment decision making in proximal humeral fractures.

After evaluation of 116 consecutive proximal humeral fractures, observers were asked to give treatment recommendation (conservative vs. surgery). If surgery was proposed, they were told to select surgery of choice.

When 3D imaging was added, complexity of fractures significantly increased ($p<0.001$), number of surgeries significantly increased ($p<0.000$) and number of ORIF treatments significantly increased ($p<0.0004$).

Addition of 3D imaging of proximal humeral fractures significantly increases number of surgical decisions when compared to radiographs alone or together with CT.

Keywords

Proximal humeral fractures; classification; surgeon decision-making; intraobserver agreement; interobserver agreement; concordance; reproducibility.

1. Introduction

Proximal humeral fractures (PHF) remain a source of controversy not only because there is a low agreement and reproducibility of the classification systems but also because there is a low agreement relative to treatment management.

Several studies have shown modest reliability and reproducibility of the classification systems for proximal humeral fractures. It seems that the poor agreement evidenced is not influenced by the classification system used (Neer's classification system, AO classification), the different imaging techniques used to characterize the fracture (plain X-ray, CT scan or 3D reconstruction) or the experience of observers participating in the study (residents, general orthopedists or upper extremity specialists) [1-6]. Only the training of the observers before the observations resulted in an improvement in agreement [7].

Even though surgeons seem to agree more on the treatment recommendations than the fracture classification, the agreement reached is only moderate [8]. Several factors have been implicated in the decision making around the operative treatment of proximal humeral fractures. They are age, fracture mechanism, associated orthopedic injuries requiring surgery and the morphological aspects of the fracture pattern [9]. Differences in agreement do not seem to be influenced by the different subspecialties of observers or either the surgeons' experience [10].

The objective of this study was to analyze the effect of the different imaging modalities in treatment decision making in proximal humeral fractures and in the agreement with the Neer's classification system [11].

2. Methods

All proximal humeral fractures from January 2012 to December 2014 were included. The exclusion criteria were isolated greater tuberosity fractures, pathological fractures, the presence of a previous PHF, fracture-dislocations and incomplete imaging (at least two plain radiographs and a CT scan).

After inclusion and exclusion criteria were applied, 116 consecutive PHF were eligible for the study.

Eight orthopedic residents in their two last years of residency were chosen as observers and were divided into two groups, the so-called addition group and the so-called sequential group. All the observers had to respond to three issues presented for every set of images. The first was to classify the fracture according to a simplified Neer's classification system, being either 1-part, 2-part, 3-part or 4-part. The second dealt with treatment decision making, conservative or surgery. The last was to choose the appropriate surgical treatment if surgery was proposed. The surgical options were open reduction internal fixation (ORIF), Hemiarthroplasty (HA) or Reverse Shoulder Arthroplasty (RSA). The responses had to be given under the supposition that all the sets of images belonged to the same healthy 72-year-old female living on her own to minimize different observer's preferences for treatment depending on age.

The addition group, on the first round, only had plain radiographs (AP and Outlet view) and had to respond to the three issues based on the plain radiographs. The observation was repeated after 15 days. In the second round, CT-scan images were added to the plain radiographs. Therefore, the responses were given after analysis of both radiographs and CT-scans. The observation was also repeated after 15 days. In the third round, 3D-reconstructed images were added to the plain radiographs and CT-scan images. In that instance, the responses were given after the analysis of the radiographs, CT scans and 3D-reconstructed images. The observation was also repeated after another subsequent 15 days.

The sequential group only had plain radiographs (AP and Outlet view) and had to respond to the three issues based on the plain radiographs in the first round. The observation was repeated after 15 days. In their second round, they only had CT-scan images. Thus, the responses were given after the analysis of CT-scans. The observation was also repeated after another 15 days. In the third round, they only had 3D-reconstructed images and the responses were given after the analysis of the 3D-reconstructed images. The observation was also repeated after then another 15 days.

In both groups, the images were prepared in PowerPoint. The order of the images was randomly changed from one observation to another.

All the observers were trained to use the Neer's classification system for PHF before the study started. The observers were not aware of the purpose or the design of the study.

2.1. Data analysis

To quantify the intra-evaluator agreement between the first and the second observation, Cohen's kappa coefficient was applied. It was calculated, together with the corresponding 95% confidence interval, using the data of all the observers. It was done separately for each question and each of the diagnostic tools available. According to the criteria of Landis and Koch, ≥ 0.81 is almost perfect agreement, between 0.61 and 0.80 is substantial agreement, between 0.41 and 0.60 is moderate agreement, between 0.21 and 0.40 is fair agreement and ≤ 0.20 is slight agreement [12]. Interobserver agreement was measured by means of the intraclass correlation coefficient based on two-way random effects models. These computations together with the 95% confidence intervals were also done for each combination of issues presented and the diagnostic tools. Added objectives of this study were to analyze whether the availability of 3D images augments the probabilities of diagnosing more complex fractures, indications for surgery and ORIF or a prosthesis, respectively. For each of these issues, generalized linear mixed models with a logic link function were used. These models included a diagnostic tool and a group (sequential versus addition) as fixed effects and an observer and image as random factors.

The statistical analyses were carried out using the statistical software program R, version 3.3.1 (Vienna, Austria; http://www.r-project.org/). In particular, the "irr" package in R[1] was used for the intra and inter-evaluator agreement analyses and the lme4 package was used to fit the generalized linear mixed models [13]. Statistical significance was set at 0.05.

The study was approved by the Ethical Committee with number 2016/6653/I (CEIC-Parc de Salut Mar).

3. Results

The mean interobserver Cohen's Kappa coefficient for the simplified Neer`s classification system, when only X-Ray was available, was κ=0.50. Analyzing by group, the addition group scored κ=0.52 and the sequential group scored κ=0.50. When the CT scan was added, κ=0.46 was the mean. Analyzing by group, the addition group scored κ=0.48 and the sequential group scored κ=0.53. When 3D was added, the mean was κ=0.46. Analyzing by group, the addition group scored κ=0.52 and the sequential group scored κ=0.46. The mean intraobserver Cohen's Kappa coefficient for the simplified Neer`s classification system when only X-Ray was available was κ=0.48. Analyzing by group, the addition group scored κ=0.48 and the sequential group scored κ=0.47. When CT scan was added, the mean was κ=0.44. Analyzing by group, the addition group scored κ=0.45 and the sequential group scored κ=0.43. When 3D was added, the mean was κ=0.51. Analyzing by groups, the addition group scored κ=0.53 and the sequential group scored κ=0.48. (Table 1)

When 3D imaging was added, the number of 3 and 4-part fractures observed significantly increased ($p<0.001$). No significant differences were noted in the number of 3 and 4-part fractures observed when the CT scan was added to X-Ray ($p=0.18$) or between the addition and the sequential groups ($p=0.37$). (Table 2)

The mean interobserver Cohen's Kappa coefficient for the treatment proposed when only X-Ray was available was κ=0.40. Analyzing by group, the addition group scored κ=0.39 and the sequential group scored κ=0.39. When CT scan was added, κ=0.33 was the mean. Analyzing by group, the addition group scored κ=0.34 and the sequential group scored κ=0.42. When 3D was added, the mean was κ= 0.29. Analyzing by groups, the additive group scored κ=0.36 and the sequential group scored κ=0.32. The mean intraobserver Cohen's Kappa coefficient for the treatment proposed when only X-Ray was available was of κ=0.55. Analyzing by groups, the additive group scored κ=0.60 and the sequential group scored κ=0.48. When CT scan was added, the mean was κ=0.52. Analyzing by groups, the additive group scored κ=0.47 and the sequential group scored κ= 0.57. When 3D was added, the mean was κ=0.58. Analyzing by groups, the additive group scored κ=0.52 and the sequential group scored κ=0.65. (Table 1)

When 3D imaging was added, the number of surgeries proposed significantly increased (p<0.000). Conversely, when the CT scan was added, the number of surgeries proposed significantly decreased (p<0.000). No significant differences were noted in the number of surgeries proposed between the addition and the sequential groups (p=0.07). (Table 2)

The mean interobserver Cohen's Kappa coefficient for the appropriate surgical treatment when only X-Ray was available was κ=0.47. Analyzing by groups, the addition group scored κ=0.48 and the sequential group scored κ=0.43. When CT scan was added, the mean was κ=0.40. Analyzing by groups, the addition group scored κ=0.44 and the sequential group scored κ=0.41. When 3D was added, the mean was κ=0.42. Analyzing by groups, the additive group scored κ=0.48 and the sequential group scored κ=0.39. The mean intraobserver Cohen's Kappa coefficient for the appropriate surgical treatment when only X-Ray was available was of κ= 0.47. Analyzing by groups, the additive group scored κ= 0.50 and the sequential group scored κ=0.43. When the CT scan was added, the mean was κ=0.49. Analyzing by groups, the additive group scored κ= 0.45 and the sequential group scored κ=0.51. When 3D was added, the mean was κ=0.52. Analyzing by groups, the addition group scored κ=0.54 and the sequential group scored κ=0.48. (Table 1)

When 3D imaging was added, the number of ORIF treatment proposed significantly increased ($p<0.0004$). No significant differences were noted in the number of ORIF proposed when the CT scan was added to X-Ray ($p=0.62$) or between the addition and the sequential groups ($p=0.37$). (Table 2)

4. Discussion

Proximal humeral fractures remain a source of controversy not only because of the lack of agreement and concordance when trying to classify them but also for the lack of consensus about the best treatment option for every individual case. In the present study, the different imaging techniques available to study PHF do not improve agreement and concordance when using Neer's classification system, as has been previously described. Moreover, when 3D reconstruction is added to the imaging studies, the appreciation of complexity in the fracture pattern is increased, and also both the number of indications for surgery and ORIF indications increase.

The lack of agreement and reproducibility of the Neer's classification system has been clearly documented. The first reports on that questioned the consistency of the classification system when assessed only with X-ray, obtaining fair to poor agreement, despite the type of images delivered or the simplification of the classification system [5,6]. Later, the addition of the CT images to assess the fracture pattern failed to improve agreement. Sjödén et al. conducted a study that included 25 proximal humeral fractures analyzed by 10 observers (5 orthopedic specialists and 5 radiology specialists) with both plain radiographs and CT. They concluded that even with the addition of the CT, the Neer's classification system showed low consistency [6]. More recent studies introducing the 3D exam of the fractures also failed to improve agreement and concordance. [1]

Agreement among doctors using the Neer's classification system does not improve through the selection of experienced observers or with advanced imaging modalities or with the simplification of the classification system [1-6]. Only the training of the observers seems to improve agreement in both experts and non-experts [7].There is no consensus as to whether agreement is better in the more or less complex fractures.

While Sjödén et al. found that agreement tended to be better in less complex fractures [6], Foroohar et al. found that advanced imaging was helpful in 4-part fracture classification [3].

In the present study, the concordance and reproducibility of the Neer's classification system was not improved by adding information through the different imaging modalities. Regardless of the way the information was delivered (additive or sequential groups), the addition of CT and 3D imaging did not improve agreement, which remained moderate. Interestingly, the total number of three and four-part fractures significantly increased when 3D imaging was introduced in both groups, reflecting that the fact 3D imaging causes an increase in the perception of the complexity of the fracture pattern. By adding CT imaging, it did not significantly change the fracture classification distribution when compared to the analysis done with X-Ray alone. Brorson et al., in a multi-centric study involving 5 observers and 193 radiographs, concluded that surgeons agree more on treatment recommendations than on classifications of proximal humeral fractures with moderate agreement relative to treatment recommendations compared to fair agreement on fracture classification. However, observers changed in 36% of their observations relative to the Neer category and in 34% of the treatment recommendation between the first and the second round [15]. Petit et al. also found a moderate agreement in terms of treatment recommendations in a study that included eight fellowship-trained orthopedic surgeons (3 shoulder, 5 trauma) who analyzed 38 radiographs. The results did not differ between subspecialties or surgeons' experience. Only by reducing the heterogeneity of responses (from 6 options of treatment to 3) did agreement increase from 0.41 to 0.48 [10].

Hageman et al. studied the factors influencing the decision making around the operative treatment of proximal humeral fractures and concluded that only the region of practice of the surgeon and delivering radiographs together with some patient information (especially age and fracture mechanism) significantly influenced decision making [14]. Okike et al. also concluded that the predictors of surgery decision in proximal humeral fractures were younger age, associated orthopedic injuries and being a shoulder or an upper extremity specialist [9].

The influence of the different imaging modalities on decision making in proximal humeral fractures has not been previously studied. The results of the present study support the hypothesis that the addition of 3D imaging significantly increases the decision to go forward with surgery (from 75.3% when only radiographs are available to 82.2% when 3D imaging is added). The addition of CT imaging has a little no significant effect on decision making when compared to radiographs alone. Moreover, 3D images also significantly increase the indication for ORIF (from 42% when only radiographs are available to 48.9% when 3D imaging is added), which reflects a current trend in the treatment of PHF. The absolute rate of surgically managed proximal humeral fractures has risen from 12.5% to 15.7%, meaning a relative increase of 25.6%, within the period from 1999 to 2000 compared to the period from 2004 to 2005. Moreover, the relative increase in the percentage of proximal humeral fractures treated with ORIF was of 28.5%. The authors wonder whether this increase in surgical treatment is due to new surgical techniques, more demanding patient expectations or the increasing skills of fellowship-trained surgeons [15]. Considering the findings of this study, newly developed imaging techniques may also have a roll in the increase in surgical indications in proximal humeral fractures. The results of the present study can be interpreted in different ways, 3D reconstruction might not be as frequently used as X-Ray and CT scan, and the observers are not familiarized with 3D images. It is also possible that the fact that the 3D images are displayed on a 2D screen affect its interpretation. Not having a “gold standard” classification system or for treatment recommendations, the authors cannot reach a conclusion as to the positive or negative effect of 3D reconstruction on the Neer`s classification system or on treatment recommendations. However, the results obtained reflect the actual trend in the PHF management and introduce imaging as a consistent factor to be considered in treatment decision making in PHF.

The limitations of the present study include the selection of observers, which may not be representative of specialized shoulder surgeons. However, it has been previously reported that the results may not differ between subspecialties or the surgeons' experience. The authors also wanted to avoid the influence that the experience that specialized shoulder surgeons have as it could influence the answers to the questions. Another limitation is to assume that all the fractures belonged to a fictional 72-year-old lady rather than providing information on every single case but the authors wanted to minimize the influence that age, gender and health status could have on decision making, leaving imaging as the only independent variable.

4. 1. Conclusions

In conclusion, the addition of 3D imaging of proximal humeral fractures significantly increases the number of surgical decisions and the number of ORIF indications when compared to radiographs alone or together with CT images. The addition of CT images does not influence decision making in proximal humeral fractures when compared to radiographs alone. The different imaging techniques do not improve the agreement or concordance of the Neer's classification system.

Financial support

None.

Financial remuneration

None.

Institutional Review Board (IRB), ethical committee approval

The study was approved by the Ethical Committee with number 2016/6653/I (CEIC-Parc de Salut Mar).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

None.

REFERENCES

1. Berkes MB, Dines JS, Little MTM, et al. The impact of three-dimensional CT imaging on intraobserver and interobserver reliability of proximal humeral fracture classifications and treatment recommendations. J Bone Joint Surg. 2014;96:1281-1286.
2. Brorson S, Bagger J, Sylvest A, et al. Low agreement among 24 doctors using the Neer-classification; only moderate agreement on displacement, even between specialists. Int Orthop. 2002;26:271-273.
3. Foroohar A, Tosti R, Richmond JM, et al. Classification and treatment of proximal humerus fractures: inter-observer reliability and agreement across imaging modalities and experience. Journal of Orthopaedic Surgery and Research. 2011;6:38.
4. Sidor ML, Zuckerman JD, Lyon T, et al. The Neer classification system for proximal humeral fractures. J Bone Joint Surg. 1993;75-A:1745-1750.
5. Siebenrock KA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. J Bone Joint Surg. 1993;75-A:1751-1755.
6. Sjödén GO, Movin T, Güntner P, et al. Poor reproducibility of classification of proximal humeral fractures: additional CT of minor value. Acta Orthop Scan. 1997;68:239-242.
7. Brorson S, Hróbjartsson A. Training improves agreement among doctors using the Neer system for proximal humeral fractures in a systematic review. Journal of Clinical Epidemiology. 2008;61:7-16.

8. Brorson S, Olsen BS, Frich LH, et al. Surgeons agree more on treatment recommendations than on classification of proximal humeral fractures. BMC Musculoskeletal Disorders. 2012;13:114.
9. Okike K, Lee OC, Makanji H, et al. Factors associated with the decision for operative versus non-operative treatment of displaced proximal humerus fractures in the elderly. Injury. 2013;44:448-453.
10. Petit CJ, Millett PJ, Endres NK, et al. Management of proximal humeral fractures: Surgeons don't agree. J Shoulder Elbow Surg. 2010;19:446-451.
11. Neer CS. Displaced proximal humeral fractures. Part I. Classification and evaluation. J Bone Joint Surg. 1970;52-A:1077-1089.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-174.
13. Bates D, Mächler M, Bolker BM, et al. Fitting linear mixed-effects models using lme4. Journal of Stastical Software. 2015;67:1-48.
14. Hageman MGJS, Jayakumar P, King JD, et al. The factors influencing the decision making of operative treatment for proximal humeral fractures. J Shoulder Elbow Surg. 2015;24:e21-e26.
15. Bell JE, Leung BC, Spratt KF, et al. Trends and variation in incidence, surgical treatment, and repeat surgery of proximal humeral fractures in the elderly. J Bone Joint Surg 2011;93:121-131.

Table I Inter- and intra-observer agreement for the three questions

	Neer's		Treatment decision (conservative/surgery)		Treatment proposal (ORIF/HA/RSA)	
	κ	95% CI	κ	95% CI	κ	95% CI
Inter-obs						
X-Ray						
All	0.50	(0.42-0.52)	0.40	(0.32-0.48)	0.47	(0.39-0.55)
Additive	0.52	(0.42-0.61)	0.39	(0.30-0.49)	0.48	(0.38-0.57)
Sequential	0.50	(0.41-0.60)	0.39	(0.29-0.50)	0.43	(0.31-0.55)
CT scan						
All	0.46	(0.37-0.55)	0.33	(0.25-0.41)	0.40	(0.32-0.49)
Additive	0.48	(0.35-0.59)	0.34	(0.24-0.45)	0.44	(0.34-0.54)
Sequential	0.53	(0.43-0.63)	0.42	(0.32-0.52)	0.41	(0.26-0.55)
3D						
All	0.46	(0.38-0.55)	0.29	(0.22-0.38)	0.42	(0.32-0.51)
Additive	0.52	(0.41-0.62)	0.36	(0.25-0.47)	0.48	(0.36-0.59)
Sequential	0.46	(0.34-0.58)	0.32	(0.22-0.43)	0.39	(0.22-0.54)
Intra-obs						
X-Ray						
All	0.48	(0.43-0.52)	0.55	(0.49-0.61)	0.47	(0.42-0.51)
Additive	0.48	(0.42-0.55)	0.60	(0.52-0.69)	0.50	(0.44-0.56)
Sequential	0.47	(0.40-0.53)	0.48	(0.39-0.58)	0.43	(0.36-0.49)
CT scan						
All	0.44	(0.40-0.49)	0.52	(0.46-0.58)	0.49	(0.45-0.54)
Additive	0.45	(0.38-0.52)	0.47	(0.38-0.55)	0.45	(0.38-0.51)
Sequential	0.43	(0.37-0.50)	0.57	(0.48-0.66)	0.51	(0.44-0.57)
3D						
All	0.51	(0.46-0.56)	0.58	(0.51-0.65)	0.52	(0.48-0.57)
Additive	0.53	(0.47-0.60)	0.52	(0.43-0.62)	0.54	(0.47-0.60)
Sequential	0.48	(0.41-0.55)	0.65	(0.55-0.75)	0.48	(0.42-0.55)

ORIF, open reduction internal fixation; HA, hemiarthroplasty; RSA, reverse shoulder arthroplasty; κ, kappa value; CI, confidence interval; Inter-obs, inter-observer; Intra-obs, intra-observer

Table II Change in the answers of the three questions when different imaging modalities were added

	Neer's 1-2-part/3- 4-part	Treatment Decision Conservative/Surgery	Treatment Proposal ORIF/HA-RSA
X-Ray			
All	44.8%/55.2%	24.7%/75.3%	42%/58%
Additive	40.7%/59.3%	27.9%/72.1%	39.1%/60.9%
sequential	48.9%/51.1%	21.4%/78.6%	44.8%/55.2%
CT scan			
All	43.1%/56.9%	28.8%/71.2%	42.2%/57.8%
Additive	39%/61%	34.8%/65.2%	33.6%/66.4%
Sequential	47.3%/52.7%	22.8%/77.2%	48.8%/51.2%
3D			
All	39%/61%	17.8%/82.2%	48.9%/51.1%
Additive	37.1%/62.9%	21.4%/78.6%	43.1%/56.9%
Sequential	40.9/59.1%	14.2%/85.8%	53.9%/46.1%
p value	<.001	<.000	.0004

ORIF, open reduction internal fixation; HA, hemiarthroplasty; RSA, reverse shoulder arthroplasty